W586–W590 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 18 May 2020*
doi: 10.1093/nar/gkaa378

InterPred: a webtool to predict chemical autofluorescence and luminescence interference

Alexandre Borrel [1], Kamel Mansouri[2], Sue Nolte[3], Trey Saddler[3], Mike Conway[3], Charles Schmitt[3] and Nicole C. Kleinstreuer [1,4,*]

[1]NIH/NIEHS/DIR/BCBB, RTP, NC 27709, USA, [2]Integrated Laboratory Systems, Inc. RTP, NC 27560, USA, [3]NIH/NIEHS/ODS, RTP, NC 27709, USA and [4]NIH/NIEHS/DNTP/NICEATM, RTP, NC 27709, USA

Received March 09, 2020; Revised April 10, 2020; Editorial Decision April 29, 2020; Accepted April 29, 2020

ABSTRACT

High-throughput screening (HTS) research programs for drug development or chemical hazard assessment are designed to screen thousands of molecules across hundreds of biological targets or pathways. Most HTS platforms use fluorescence and luminescence technologies, representing more than 70% of the assays in the US Tox21 research consortium. These technologies are subject to interferent signals largely explained by chemicals interacting with light spectrum. This phenomenon results in up to 5–10% of false positive results, depending on the chemical library used. Here, we present the InterPred webserver (version 1.0), a platform to predict such interference chemicals based on the first large-scale chemical screening effort to directly characterize chemical-assay interference, using assays in the Tox21 portfolio specifically designed to measure autofluorescence and luciferase inhibition. InterPred combines 17 quantitative structure activity relationship (QSAR) models built using optimized machine learning techniques and allows users to predict the probability that a new chemical will interfere with different combinations of cellular and technology conditions. InterPred models have been applied to the entire Distributed Structure-Searchable Toxicity (DSSTox) Database (∼800,000 chemicals). The InterPred webserver is available at https://sandbox.ntp.niehs.nih.gov/interferences/.

INTRODUCTION

Chemical hazard assessment screening and drug discovery programs use an array of cell-based assays measuring processes such as cell growth/death, receptor binding, or protein expression, while others rely upon cell-free assays that characterize biochemical activity. Both formats use mostly fluorescence-based detection technologies as showed in applications reported in PubChem (1,2). This type of technology allows for optimization of speed, accuracy, reproducibility and assay sensitivity (3). Specifically, luminescence technology is frequently used as a readout from luciferase-based reporter gene assays and provides high sensitivity due to lack of background activity in mammalian cell lines.

These types of technology are prone to interference by chemicals that may modulate the signal intensity without displaying any true biological activity. Chemicals may interfere with fluorescent assays via quenching, direct absorption of light, or autofluorescence, where they emit light that overlaps with the fluorophore spectrum (4). Luciferase assays are subject to interference via direct inhibition of enzymatic activity or oxidation of the luciferin substrate (5). This is a fairly common phenomenon, where autofluorescence has been observed for >5% of PubChem chemical libraries (6), and unexplained luminescence changes occurred in 12% of active chemicals from the NIH Molecular Libraries Small Molecule Repository (7).

The first large-scale screening effort to directly characterize chemical-assay interference, rather than as an observed byproduct of measuring biological activity, used multiple assays in the Tox21 portfolio (8,9) specifically designed to measure autofluorescence and luciferase inhibition. Over 8000 unique structures covering environmental toxicants of regulatory concern, food additives, pharmaceuticals, and industrial chemicals, many with significant human exposure potential, were run using ultra high-throughput screening (HTS) technologies measuring interference with luciferase- and fluorescence-based readouts under various culture conditions and cell types. Advanced cheminformatics approaches were used to relate chemical structural clusters to interference activity profiles, and multiple machine learning algorithms were applied to predict assay interference based on molecular descriptors and physicochemical properties (10).

Here, we present InterPred, a webserver hosted by the National Toxicology Program (https://sandbox.ntp.niehs.nih.gov/interferences/) including the best performing inter-

*To whom correspondence should be addressed. Tel: +1 984 287 3150; Email: nicole.kleinstreuer@nih.gov

ference predictive models (accuracies of ∼80%). The InterPred tool allows users to predict the likelihood of assay interference for any new chemical structure, increasing confidence in HTS data by decreasing false positive testing results.

MATERIALS AND METHODS

Tox21 interference library

Three assay platforms from the Tox21 portfolio were used to develop interference QSAR models (10). The raw data are freely available on the NCATS Tox21 browser (https://tripod.nih.gov/tox21/assays/) under the names 'tox21-luc-biochem-p1' for the luciferase inhibition assay, and 'tox21-spec-hepg2-p1' and 'tox21-spec-hek293-p1' for autofluorescence assays using HepG2 and HEK-293 cell cultures, respectively, measuring red, green and blue wavelengths using cell-based and cell-free culture-medium-only conditions. The Tox21 chemical library (8,305 unique substances) was screened in triplicate concentration response in all assays, with concurrent cytotoxicity measurements where applicable.

Molecular modeling

Chemicals are encoded using a unique SMILES string format. Using the MolVS python library available at http://molvs.readthedocs.io/en/latest/guide/intro.html each chemical is prepared from original SMILES including the following steps: hydrogen removing, sanitization, metal disconnection, stereochemistry process, desolvation and filtering of salt fragments. Mixtures were not considered and were excluded in an early step. From each curated structure, a set of 677 1D–2D descriptors is computed using RDKit tool kit version 2019-09-03 (https://www.rdkit.org/) and an additional set of 30 physicochemical descriptors is computed using the OPERA models version 2.3 (11).

InterPred QSAR models

Data chemical curation descriptor selection and QSAR modelling workflows were conducted according to best practices (12–15). Classification models to predict active versus inactive chemicals for each of the interference assay endpoints were built using four machine learning approaches, see (10) for details. Considering the unbalanced dataset, i.e. far more inactive chemicals as shown in Supporting Information Table S1, under-sampling methods were applied via random selection of inactive chemicals to yield a ratio of 70% inactive and 30% active chemicals. All model building steps were repeated 10 times to ensure that all chemicals were incorporated in the process. Only the best performing QSAR models, developed using random forest machine learning (16), were selected for inclusion in the InterPred webserver. For every model, loss of performance can be explained by low sensitivity, i.e. the difficulty in predicting active chemicals. This phenomenon is explained by the structural diversity of the large inactive set, and the coverage of the active set by the inactive set. Models were validated using 10-fold cross validation and an external test set. Performances in cross validation and on an independent test set are presented in Supporting information Tables S2 and S3. The InterPred webserver allows users to predict chemical interference using 17 QSAR models including: luminescence/luciferase, overall autofluorescence, autofluorescence on the blue, green and red wave lengths individually, and all combinations of autofluorescence interference for assays using HepG2 or HEK-293 cell lines in cell free or cell-based conditions, with three wave lengths each (blue, green and red). Each model output is an interference probability (from 0 to 1) for each chemical/assay combination with a standard deviation correlated to the model confidence, see methods in (10). Scripts developed to build InterPred QSAR models are freely available on a GitHub directory (https://github.com/ABorrel/interferences), and the modeling dataset is included in an archive in Supporting Information (File S1).

Distributed structure-searchable toxicity (DSSTox) database prediction

All 17 models were run to produce interference predictions for all structures included in the Distributed Structure-Searchable Toxicity (DSSTox) Database (https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database) containing ∼800 000 chemicals (17,18). The coverage of the applicability domain of the Tox21 chemical library on the DSSTox DB is discussed in (10) and the broad coverage of the structural landscape of Tox21 chemicals on the principal component analysis defined from the DSSTOX chemical library using 1D and 2D molecular descriptors is presented in Supporting information Figure S1.

Webserver development

The webserver was developed using Django in Python 3.6.8 on a CentOS 7 virtual machine. Interference prediction models were developed in R 3.4.4. Interactive result tables were developed using the agGrid (https://www.ag-grid.com/) JavaScript library. InterPred webserver includes a PostgreSQL database used to store molecular descriptors, interference prediction probabilities and prepared structures. More than 800 000 chemical entries are included in the database.

RESULTS

Workflow

The InterPred protocol is presented as a workflow in Figure 1.

Input. Users can upload up to 100 chemicals on the server in a SMILES string format or using CASRN or DSSTox identifier. Chemicals can be pasted in the text bar or uploaded in a text format with one chemical per line. In the first step, InterPred verifies each entry and transforms CASRN and DSSTox identifiers into a SMILES string format.

Chemical preparation. Following the best practices, see methods, each chemical structure is prepared and cleaned. At the end of this step users can decide to resubmit

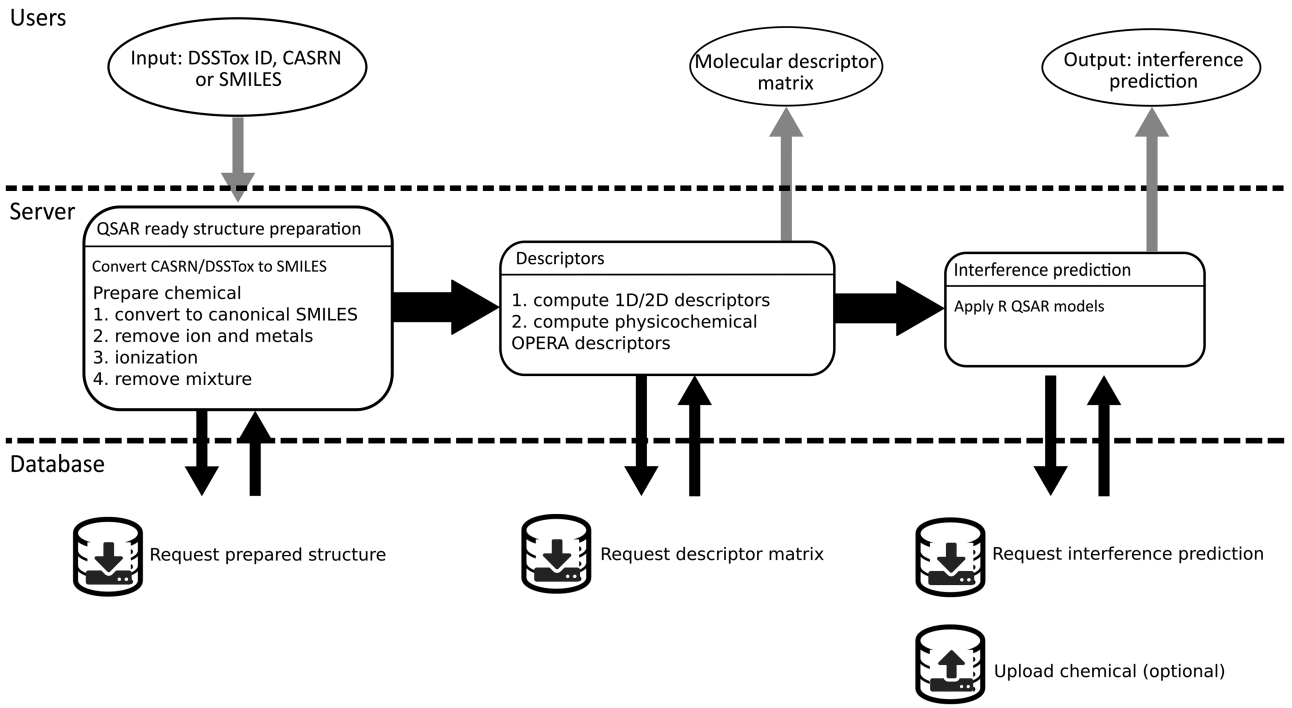

Figure 1. Workflow presenting the different steps of the InterPred chemical-assay interference prediction protocol. The top portion represents user options, input and output, in the middle the different server steps, and in the bottom the database interactions.

Table 1. Percentage of interference chemicals predicted from the DSSTox database, by technology (luciferase inhibition, autofluorescence) and by cell culture condition. Percentage of predicted active chemicals is reported for different probability cutoffs

			% Interferent chemicals				
QSAR model	Cell culture	Conditions/endpoint	Cutoff > 0.5	Cutoff > 0.6	Cutoff > 0.7	Cutoff > 0.8	Cutoff > 0.9
Luciferase		Luciferase	18.21	8.83	3.09	0.64	0.04
Auto-fluorescence	HepG2 HEK-293	All	10.96	5.71	2.13	0.25	0.02
		Blue	8.47	3.07	0.65	0.04	0
		Green	5.78	2.16	0.5	0.09	0.01
		Red	2.95	0.97	0.34	0.06	0.01
	HepG2	Cell based blue	11.57	6.09	2.69	0.52	0.01
		Cell based green	5.81	2.48	0.78	0.17	0.02
		Cell based red	4.67	1.55	0.4	0.1	0.01
		Cell free blue	12.21	5.64	2	0.21	0
		Cell free green	4.36	1.54	0.52	0.14	0.02
		Cell free red	3.64	1.28	0.45	0.1	0.01
	HEK-293	Cell based blue	11.96	5.21	1.78	0.2	0
		Cell based green	4.4	1.32	0.36	0.07	0.01
		Cell based red	3.54	1.21	0.34	0.07	0.01
		Cell free blue	10.18	4.22	1.1	0.05	0
		Cell free green	3.01	0.69	0.15	0.04	0.01
		Cell free red	3.88	1.54	0.58	0.2	0.03

their chemicals or continue the protocol. Chemicals in the DSSTox database are already prepared and included in the database. To simplify the database requests, cleaned chemicals in SMILES format are converted into InChIKey format.

Molecular descriptors. Molecular descriptors are computed on the fly using custom Python scripts. The database was populated with the full descriptor set for ∼800k chemicals in the DSSTox database. At this step, for their list of chemicals, users can download the 1D/2D RDKit molecular descriptor matrix as well as physicochemical descriptor matrix computed using OPERA models.

Interference prediction. Users can choose up to 17 interference QSAR models to apply on the input chemical list. For each chemical, each model returns a probability (from 0 to 1, with an associated standard deviation) that that chemical will interfere with the particular assay technology (i.e. luciferase inhibition or autofluorescence under various wavelengths and cell culture conditions). The results page includes a dynamic table where users can sort chemicals using

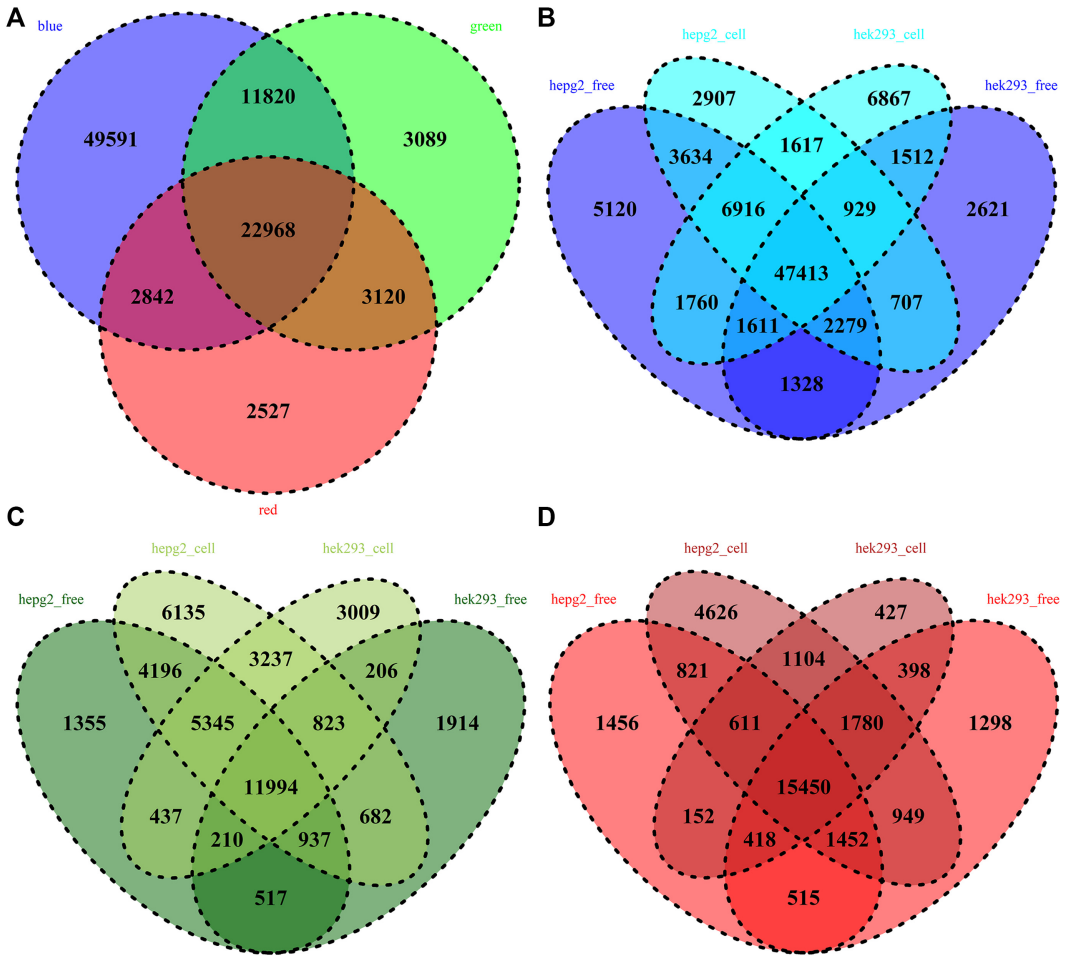

Figure 2. Venn diagrams of predicted interferent chemicals from the ∼800 000 compound DSSTox library, with a probability cutoff of 0.5, (**A**) comparing different color channels, (**B**) considering only the blue channel and comparing cell lines/culture conditions, (**C**) considering only the green channel and comparing cell lines/culture conditions and (**D**) considering only the red channel and comparing cell lines/culture conditions. For specific color channels cell culture conditions are represented as cell based (hepg2_cell and hek293_cell) and cell free, culture medium only (hepg2_free and hek293_free).

the interference probability on a specific selected model. In addition, all predictions are downloadable in a .csv format.

Computational time and database storage. Computational processing time is dependent on molecular size and complexity. For example, a small molecule with less than ten atoms can be processed in few seconds while a chemical with >40 atoms will take more than ten seconds. To reduce waiting time, we limited the number of chemicals that can be uploaded by the user to one hundred. The entire DSSTox database was precomputed, covering >800 000 chemical structures. In addition, by default we store any new chemical uploaded on the webserver in the database. This approach speeds up the interference prediction by limiting duplicated server runs. However, being aware that users' data can be sensitive or confidential, we allow users to opt out of saving chemicals in the database.

DSSTox interference predictions

The interference prediction workflow contained in InterPred was applied to the DSSTox database (https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database) including >800 000 chemicals. After the first chemical preparation step, 573 841 structures in QSAR ready form were used to compute descriptors and predict assay technology interference with the 17 QSAR models. For each chemical, each QSAR model produced a probability that the chemical would interfere with that specific assay technology/condition. Table 1 summarizes the number of interferent chemicals found for each model with different probability cutoffs. The number of interferent chemicals was correlated with the probability cutoff, which users can specify. Only a few chemicals were identified as interferent with a probability >0.9. Globally, the same tendency was found as in the training set, see (10), with the largest number of interferent chemicals predicted for luciferase inhibition and autofluorescence in the blue channel. The coverage of the number of predicted interferent chemicals across different cell types, conditions and colors, is presented in Figure 2. As expected, significant portions of the chemicals were predicted interferent in several conditions by color. As discussed in (10), different types of interferent chemicals

can be identified by an analysis combining QSAR model results. For example, a chemical that interferes on all of the same color conditions is most likely to interfere with the light spectrum. Analogously, a chemical predicted to be active across models in the same cell type is more likely to interfere with the cell, e.g. by interfering with cell metabolism. From the DSSTox predictions, the top 8 chemicals predicted active across all models are presented in Supporting Information Figure S2. As expected, these chemicals contained a complex aromatic ring arrangement composed of more than 3 rings which correspond to light absorption property, and included dye chemicals such as Erythrosin (15905-32-5) or reference chemicals such as Fluorescein (2321-07-5) or Rose Bengal (152-75-0).

DISCUSSION

Here, we have presented an intuitive, easy to use webserver combining 17 models predicting the likelihood of chemical interference with the most commonly used assay technologies in large screening efforts. The InterPred website is free and open to all users and there is no login requirement. The database has been populated with interference predictions on over 800 000 chemicals. This work builds on previous efforts to predict chemical-assay interference using rule-based classification models (19) and substructure filters (20) by applying machine learning to a large dataset of chemicals screened specifically for assay interference across technology types and cell culture conditions. By providing a range of QSAR models and a probabilistic prediction approach, we allow users to assemble predictions on different technologies and conditions to build a better understanding of the potential interferent properties of a new chemical. Notably, this webserver can also be used to compute 1D/2D chemical descriptors as well as predict physicochemical properties from OPERA models. We feel that this work is an extremely valuable contribution to the scientific community by providing novel insight into structural patterns driving false signals in the most common HTS assays, and therefore has the potential to save significant time and resources in both the study design and data analysis phases.

DATA AVAILABILITY

The raw data used to build QSAR models are freely available on the NCATS Tox21 browser at https://tripod.nih.gov/tox21/assays/ under the names 'tox21-luc-biochem-p1' for the luciferase inhibition assay, and 'tox21-spec-hepg2-p1' and 'tox21-spec-hek293-p1'. DSSTox chemicals are freely available on the EPA chemical dashboard at https://comptox.epa.gov/dashboard. The InterPred website (https://sandbox.ntp.niehs.nih.gov/interferences/) is free and open to all users and there is no login requirement.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: National Institute of Environmental Health Science.
Conflict of interest statement. None declared.

REFERENCES

1. Inglese,J., Johnson,R.L., Simeonov,A., Xia,M., Zheng,W., Austin,C.P. and Auld,D.S. (2007) High-throughput screening assays for the identification of chemical probes. *Nat. Chem. Biol.*, **3**, 466–479.
2. Thorne,N., Inglese,J. and Auld,D.S. (2010) Illuminating insights into firefly luciferase and other bioluminescent reporters used in chemical biology. *Chem. Biol.*, **17**, 646–657.
3. Fan,F. and Wood,K.V. (2007) Bioluminescent assays for high-throughput screening. *Assay Drug Dev. Technol.*, **5**, 127–136.
4. Sittampalam,G., Coussens,N., Arkin,M., Auld,D., Austin,C., Bejcek,B., Glicksman,M., Inglese,J., Iversen,P., Mcgee,J. *et al.* (2018) *Assay Guidance Manual Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004-*, Bethesda.
5. Auld,D.S. and Inglese,J. (2004) Interferences with Luciferase Reporter Enzymes. In: Sittampalam,GS, Grossman,A, Brimacombe,K, Arkin,M, Auld,D, Austin,CP, Baell,J, Bejcek,B, Caaveiro,JMM and Chung,TDY *et al.*Assay Guidance Manual [Internet]. Bethesda.
6. Simeonov,A., Jadhav,A., Thomas,C.J., Wang,Y., Huang,R., Southall,N.T., Shinn,P., Smith,J., Austin,C.P., Auld,D.S. *et al.* (2008) Fluorescence spectroscopic profiling of compound libraries. *J. Med. Chem.*, **51**, 2363–2371.
7. Thorne,N., Shen,M., Lea,W.A., Simeonov,A., Lovell,S., Auld,D.S. and Inglese,J. (2012) Firefly luciferase in chemical biology: a compendium of inhibitors, mechanistic evaluation of chemotypes, and suggested use as a reporter. *Chem. Biol.*, **19**, 1060–1072.
8. Collins,F.S., Gray,G.M. and Bucher,J.R. (2008) Toxicology. Transforming environmental health protection. *Science*, **319**, 906–907.
9. Thomas,R.S., Paules,R.S., Simeonov,A., Fitzpatrick,S.C., Crofton,K.M., Casey,W.M. and Mendrick,D.L. (2018) The us federal Tox21 program: a strategic and operational plan for continued leadership. *ALTEX*, **35**, 163–168.
10. Borrel,A., Huang,R., Sakamuru,S., Xia,M., Simeonov,A., Mansouri,K., Houck,K.A., Judson,R.S. and Kleinstreuer,N.C. (2020) High-throughput screening to predict chemical-assay interference. *Sci. Rep.*, **10**, 3986.
11. Mansouri,K., Grulke,C.M., Judson,R.S. and Williams,A.J. (2018) OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminform.*, **10**, 10.
12. Tropsha,A. (2010) Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.*, **29**, 476–488.
13. Golbraikh,A., Muratov,E., Fourches,D. and Tropsha,A. (2014) Data set modelability by QSAR. *J. Chem. Inf. Model.*, **54**, 1–4.
14. Cherkasov,A., Muratov,E.N., Fourches,D., Varnek,A., Baskin,I.I., Cronin,M., Dearden,J., Gramatica,P., Martin,Y.C., Todeschini,R. *et al.* (2014) QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.*, **57**, 4977–5010.
15. Fourches,D., Muratov,E. and Tropsha,A. (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.*, **50**, 1189–1204.
16. Breiman,L. (2001) Random forest. *Mach. Learn.*, **45**, 5–32.
17. Richard,A.M. and Williams,C.R. (2002) Distributed structure-searchable toxicity (DSSTox) public database network: a proposal. *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, **499**, 27–52.
18. Williams,A.J., Grulke,C.M., Edwards,J., McEachran,A.D., Mansouri,K., Baker,N.C., Patlewicz,G., Shah,I., Wambaugh,J.F., Judson,R.S. *et al.* (2017) The comptox chemistry dashboard: a community data resource for environmental chemistry. *J. Cheminform.*, **9**, 61.
19. Su,B.-H., Tu,Y.-S., Lin,O.A., Harn,Y.-C., Shen,M.-Y. and Tseng,Y.J. (2015) Rule-based classification models of molecular autofluorescence. *J. Chem. Inf. Model.*, **55**, 434–445.
20. Baell,J.B. and Holloway,G.A. (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.*, **53**, 2719–2740.